Word Salad: Relating Food Prices and Descriptions

Victor Chahuneau  
*Carnegie Mellon University*

Kevin Gimpel  
*Carnegie Mellon University*

Bryan R. Routledge  
*Carnegie Mellon University, routledge@cmu.edu*

Lily Scherlis  
*Phillips Academy*

Noah A. Smith  
*Carnegie Mellon University, nasmith@cs.cmu.edu*

Follow this and additional works at: [http://repository.cmu.edu/lti](http://repository.cmu.edu/lti)  
Part of the [Computer Sciences Commons](http://repository.cmu.edu/lti)

Published In  
Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL), 1357-1367.

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Language Technologies Institute by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
Abstract

We investigate the use of language in food writing, specifically on restaurant menus and in customer reviews. Our approach is to build predictive models of concrete external variables, such as restaurant menu prices. We make use of a dataset of menus and customer reviews for thousands of restaurants in several U.S. cities. By focusing on prediction tasks and doing our analysis at scale, our methodology allows quantitative, objective measurements of the words and phrases used to describe food in restaurants. We also explore interactions in language use between menu prices and sentiment as expressed in user reviews.

1 Introduction

What words might a menu writer use to justify the high price of a steak? How does describing an item as chargrilled vs. charbroiled affect its price? When a customer writes an unfavorable review of a restaurant, how is her word choice affected by the restaurant’s prices? In this paper, we explore questions like these that relate restaurant menus, prices, and customer sentiment. Our goal is to understand how language is used in the food domain, and we direct our investigation using external variables such as restaurant menu prices.

We build on a thread of NLP research that seeks linguistic understanding by predicting real-world quantities from text data. Recent examples include prediction of stock volatility (Kogan et al., 2009) and movie revenues (Joshi et al., 2010). There, prediction tasks were used for quantitative evaluation and objective model comparison, while analysis of learned models gave insight about the social process behind the data.

We echo this pattern here as we turn our attention to language use on restaurant menus and in user restaurant reviews. We use data from a large corpus of restaurant menus and reviews crawled from the web and formulate several prediction tasks. In addition to predicting menu prices, we also consider predicting sentiment along with price.

The relationship between language and sentiment is an active area of investigation (Pang and Lee, 2008). Much of this research has focused on customer-written reviews of goods and services, and perspectives have been gained on how sentiment is expressed in this type of informal text. In addition to sentiment, however, other variables are reflected in a reviewer’s choice of words, such as the price of the item under consideration. In this paper, we take a step toward joint modeling of multiple variables in review text, exploring connections between price and sentiment in restaurant reviews.

Hence this paper contributes an exploratory data
analysis of language used to describe food (by its purveyors and by its consumers). While our primary goal is to understand the language used in our corpus, our findings bear relevance to economics and hospitality research as well. This paper is a step on the way to the eventual goal of using linguistic analysis to understand social phenomena like sales and consumption.

2 Related Work

There are several areas of related work scattered throughout linguistics, NLP, hospitality research, and economics.

Freedman and Jurafsky (2011) studied the use of language in food advertising, specifically the words on potato chip bags. They argued that, due to the ubiquity of food writing across cultures, ethnic groups, and social classes, studying the use of language for describing food can provide perspective on how different socioeconomic groups self-identify using language and how they are linguistically targeted. In particular, they showed that price affects how “authenticity” is realized in marketing language, a point we return to in §5. This is an example of how price can affect how an underlying variable is expressed in language. Among other explorations in this paper, we consider how price interacts with expression of sentiment in user reviews of restaurants.

As mentioned above, our work is related to research in predicting real-world quantities using text data (Koppel and Shtrimerberg, 2006; Ghose et al., 2007; Lerman et al., 2008; Kogan et al., 2009; Joshi et al., 2010; Eisenstein et al., 2010; Eisenstein et al., 2011; Yogatama et al., 2011). Like much of this prior work, we aim to learn how language is used in a specific context while building models that achieve competitive performance on a quantitative prediction task.

Along these lines, there is recent interest in exploring the relationship between product sales and user-generated text, particularly online product reviews. For example, Ghose and Ipeirotis (2011) studied the sales impact of particular properties of review text, such as readability, the presence of spelling errors, and the balance between subjective and objective statements. Archak et al. (2011) had a similar goal but decomposed user reviews into parts describing particular aspects of the product being reviewed (Hu and Liu, 2004). Our paper differs from price modeling based on product reviews in several ways. We consider a large set of weakly-related products instead of a homogeneous selection of a few products, and the reviews in our dataset are not product-centered but rather describe the overall experience of visiting a restaurant. Consequently, menu items are not always mentioned in reviews and rarely appear with their exact names. This makes it difficult to directly use review features in a pricing model for individual menu items.

Menu planning and pricing has been studied for many years by the culinary and hospitality research community (Kasavana and Smith, 1982; Kelly et al., 1994), often including recommendations for writing menu item descriptions (Miller and Pavesic, 1996; McVety et al., 2008). Their guidelines frequently include example menus from successful restaurants, but typically do not use large corpora of menus or automated analysis, as we do here. Other work focused more specifically on particular aspects of the language used on menus, such as the study by Zwicky and Zwicky (1980), who made linguistic observations through manual analysis of a corpus of 200 menus.

Relatedly, Wansink et al. (2001; 2005) showed that the way that menu items are described affects customers’ perceptions and purchasing behavior. When menu items are described evocatively, customers choose them more often and report higher satisfaction with quality and value, as compared to when they are given the same items described with conventional names. Wansink et al. did not use a corpus, but rather conducted a small-scale experiment in a working cafeteria with customers and collected surveys to analyze consumer reaction. While our goals are related, our experimental approach is different, as we use automated analysis of thousands of restaurant menus and rely on a set of one million reviews as a surrogate for observing customer behavior.

Finally, the connection between products and prices is also a central issue in economics. However, the stunning heterogeneity in products makes empirical work challenging. For example, there are over 50,000 menu items in New York that include
the word *chicken*. What is the price of chicken? This is an important practical and daunting matter when measuring inflation (e.g., Consumer Price Index is measured with a precisely-defined basket of goods). Price dispersion across goods and the variation of the goods is an important area of industrial organization economic theory. For example, economists are interested in models of search, add-on pricing, and obfuscation (Baye et al., 2006; Ellison, 2005).

## 3 Data

We crawled Allmenus.com ([www.allmenus.com](http://www.allmenus.com)) to gather menus for restaurants in seven U.S. cities: Boston, Chicago, Los Angeles, New York, Philadelphia, San Francisco, and Washington, D.C. Each menu includes a list of item names with optional text descriptions and prices. Most Allmenus restaurant pages contain a link to the corresponding page on Yelp ([www.yelp.com](http://www.yelp.com)) with metadata and user reviews for the restaurant, which we also collected.

The metadata consist of many fields for each restaurant, which can be divided into three categories: location (city, neighborhood, transit stop), services available (take-out, delivery, wifi, parking, etc.), and ambience (good for groups, noise level, attire, etc.). Also, the category of food and a price range ($ to $$$$), indicating the price of a typical meal at the restaurant) are indicated. The user reviews include a star rating on a scale of 1 to 5.

The distribution of prices of individual menu items is highly skewed, with a mean of $9.22 but a median of $6.95. On average, a restaurant has 73 items on its menu with a median price of $8.69 and 119 Yelp reviews with a median rating of 3.55 stars. The star rating and price range distributions are shown in Figure 1.

The set of restaurants was randomly split into three parts (80% for training, 10% for development, 10% for evaluation), independently for each city. The sizes of the splits and the full set of dataset statistics are provided in Table 1.

### 4 Predictive Tasks

We consider several prediction tasks using the dataset just described. These include predicting individual menu item prices (§5), predicting the price range for each restaurant (§6), and finally jointly predicting median price and sentiment for each restaurant (§7). To do this, we use two types of models: linear regression (§5 and §6) and logistic regression (§7), both with \( \ell_1 \) regularization when sparsity is desirable. We tune the regularization coefficient by choosing the value that minimizes development set loss (mean squared error and log loss, respectively).

For evaluation, we use mean absolute error (MAE) and mean relative error (MRE). Given a dataset \( \langle x_i, y_i \rangle_{i=1}^N \) with inputs \( x_i \) and outputs \( y_i \), and

<table>
<thead>
<tr>
<th>City</th>
<th># Restaurants</th>
<th># Menu Items</th>
<th># Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train dev. test</td>
<td>train dev. test</td>
<td>train dev. test</td>
</tr>
<tr>
<td>Boston</td>
<td>930 107 113</td>
<td>63,422 8,426 8,409</td>
<td>80,309 10,976 11,511</td>
</tr>
<tr>
<td>Chicago</td>
<td>804 98 100</td>
<td>51,480 6,633 6,939</td>
<td>73,251 9,582 10,965</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>624 80 68</td>
<td>17,980 2,938 1,592</td>
<td>75,455 13,227 5,716</td>
</tr>
<tr>
<td>New York</td>
<td>3,965 473 499</td>
<td>365,518 42,315 45,728</td>
<td>326,801 35,529 37,795</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>1,015 129 117</td>
<td>83,818 11,777 9,295</td>
<td>52,275 7,347 5,790</td>
</tr>
<tr>
<td>San Francisco</td>
<td>1,908 255 234</td>
<td>103,954 12,871 12,510</td>
<td>499,984 59,378 67,010</td>
</tr>
<tr>
<td>Washington, D.C.</td>
<td>773 110 121</td>
<td>47,188 5,957 7,224</td>
<td>71,179 11,852 14,129</td>
</tr>
<tr>
<td>Total</td>
<td>10,019 1,252 1,252</td>
<td>733,360 90,917 91,697</td>
<td>1,179,254 147,891 152,916</td>
</tr>
</tbody>
</table>

Table 1: Dataset statistics.
denoting predicted outputs by \( \hat{y}_i \), these are defined as follows:

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|
\]

\[
\text{MRE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i}
\]

In practice, since we model log-prices but evaluate on real prices, the final prediction is often a non-linear transformation of the output of the linear classifier of weight vector \( w \), which we denote by:

\[
\hat{y}_i = f(w^\top x_i).
\]

We also frequently report the total number of features available in the training set for each model (nf) as well as the number of non-zero feature weights following learning (nnz).

5 Menu Item Price Prediction

We first consider the problem of predicting the price of each item on a menu. In this case, every instance \( x_i \) corresponds to a single item in the menu parametrized by the features detailed below and \( y_i \) is the item’s price. In this section, our models always use the logarithm of the price\(^1\) as output values and therefore:

\[
\hat{y}_i = e^{w^\top x_i}.
\]

Baselines We evaluate several baselines which make independent predictions for each distinct item name. The first two predict the mean or the median of the prices in the training set for a given item name, and use the overall price mean or median when a name is missing in the training set. The third baseline is an \( \ell_1 \)-regularized linear regression model trained with a single binary feature for each item name in the training data. These baselines are shown as the first three rows in Table 2.

We note that there is a wide variation of menu item names in the dataset, with more than 400,000 distinct names. Although we address this issue later by introducing local text features, we also performed simple normalization of the item names for all of the baselines described above. To do this normalization, we first compiled a stop word list based on the most frequent words in the item names.\(^2\) We removed stop words and then ordered the words in each item name lexicographically, in order to collapse together items such as coffee black and black coffee. This normalization reduced the unique item name count by 40%, strengthening the baselines.

5.1 Features

We use \( \ell_1 \)-regularized linear regression for feature-rich models. We now introduce several sets of features that we add to the normalized item names:\(^3\)

I. METADATA: Binary features for each restaurant metadata field mentioned above, excluding price range. A separate binary feature is included for each unique \( \langle \text{field}, \text{value} \rangle \) tuple.

II. MENUNAMES: \( n \)-grams in menu item names. We used binary features for unique unigrams, bigrams, and trigrams. Here, stop words were retained as they can be informative (e.g., with and large correlate with price).

III. MENUDESC: \( n \)-grams in menu item descriptions, as in MENUNAMES.

Review Features In addition to these features, we consider leveraging the large amount of text present in user reviews to improve predictions. We attempted to find mentions of menu items in the reviews and to include features extracted from the surrounding words in the model. Perfect item mentions being relatively rare, we consider inexact matches weighted by a coefficient measuring the degree of resemblance: we used the Dice similarity between the set of words in the sentence and in the item name. We then extracted \( n \)-gram features from this sentence, and tried several ways to use them for price prediction.

Given a review sentence, one option is to add the corresponding features to every item matching this sentence, with a value equal to the similarity coefficient. Another option is to select the best matching item and use the same real-valued features but only for this single item. Binary feature values can be used instead of the real-valued similarity coefficient. We also experimented with the use of part-of-speech tags in order to restrict our features to adjective and adverb \( n \)-grams instead of the full vocabulary. All of

---

\(^1\)The price distribution is more symmetric in the log domain.

\(^2\)This list can be found in the supplementary material.

\(^3\)The normalized item names are present as binary features in all of our regression models.
these attempts yielded negative or only slightly positive results, of which we include only one example in our experiments: the MENTIONS feature set consists of n-grams for the best matching item with the Dice coefficient as the feature value.

We also tried to incorporate the reviews by using them in aggregate via predictions from a separate model; we found this approach to work better than the methods described above which all use features from the reviews directly in the regression model. In particular, we use the review features in a separate model that we will describe below (§6) to predict the price range of each restaurant. The model uses unigrams, bigrams, and trigrams extracted from the reviews. We use the estimated price range (which we denote PR) as a single additional real-valued feature for individual item price prediction.

### 5.2 Results

Our results are shown in Table 2. We achieve a final reduction of 50 cents in MAE and nearly 10% in MRE compared with the baselines. Using menu name features (MENUNAMES) brings the bulk of the improvement, though menu description features (MENUDESC) and the remaining features also lead to small gains. Interestingly, as the MENUDESC and PR features are added to the model, the regularization favors more general features by selecting fewer and fewer non-zero weights.

While METADATA features improve over the baselines when used alone, they do not lead to improved performance over the MENU* + PR features, suggesting that the text features may be able to substitute for the information in the metadata, at least for prediction of individual item prices.

The MENTIONS features resulted in a small improvement in MAE and MRE, but at the cost of expanding the model size significantly. A look at the learned feature weights reveals that most of the selected features seem more coincidental than generic (rachel’s, highly negative) when not totally unintuitive (those suicide, highest positive). This suggests that our method of extracting features from mentions is being hampered by noise. We suspect that these features could be more effective with a better method of linking menu items to mentions in review text.

### 5.3 Analysis

We also inspected the feature weights of our learned models. By comparing the weights of related features, we can see the relative differences in terms of contribution to menu item prices. Table 3 shows example feature weights, manually arranged into several categories (taken from the model with MENUNAMES + MENUDESC + PR + METADATA).

Table 3(a) shows selected features for the “ambience” field in the Yelp restaurant metadata and pane (b) lists some unigrams related to cooking methods. Pane (c) shows feature weights for n-grams often used to market menu items; we see larger weights for words targeting those who want to eat organically- or locally-grown food (farmhouse, heirloom, wild, and hormone), compared to those looking for comfort food (old time favorite, traditional, real, and fashioned). This is related to observations made by Freedman and Jurafsky (2011) that cheaper food is marketed by appealing to tradition and historicity, with more expensive food described in terms of naturalness, quality of ingredients, and the preparation process (e.g., hand picked, wild caught, etc.). Relatedly, in pane (e) we see that real mashed potatoes are expected to be cheaper than those described as creamy or smooth.

Pane (d) shows feature weights for trigrams containing units of chicken; we can see an ordering in terms of size (bits < cubes < strips < cuts) as well as the price increase associated with the use of the word morsels in place of less refined units. We also see a difference between pieces and pcs, with the latter being frequently used to refer to entire cuts of chicken.
Table 3: Selected features from model for menu item price prediction. See text for details.

<table>
<thead>
<tr>
<th>(a) METADATA: ambience</th>
<th>(c) MENUDESC: descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>dive-y</td>
<td>old time favorite</td>
</tr>
<tr>
<td>intimate</td>
<td>fashionable</td>
</tr>
<tr>
<td>trendy</td>
<td>line caught</td>
</tr>
<tr>
<td>casual</td>
<td>all natural</td>
</tr>
<tr>
<td>romantic</td>
<td>traditional</td>
</tr>
<tr>
<td>classy</td>
<td>natural</td>
</tr>
<tr>
<td>touristy</td>
<td>classic</td>
</tr>
<tr>
<td>upscale</td>
<td>free range</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) MENUDESC: cooking</th>
<th>(e) MENUDESC: potatoes</th>
</tr>
</thead>
<tbody>
<tr>
<td>panfried</td>
<td>real mashed</td>
</tr>
<tr>
<td>chargrilled</td>
<td>mashed</td>
</tr>
<tr>
<td>boiled</td>
<td>creamy mashed</td>
</tr>
<tr>
<td>fried</td>
<td>smoked</td>
</tr>
<tr>
<td>steamed</td>
<td>locally</td>
</tr>
<tr>
<td>charbroiled</td>
<td>natural grass fed</td>
</tr>
<tr>
<td>grilled</td>
<td>Artisanal</td>
</tr>
<tr>
<td>simmered</td>
<td>raised</td>
</tr>
<tr>
<td>roasted</td>
<td>heirloom</td>
</tr>
<tr>
<td>sauteed</td>
<td>wild</td>
</tr>
<tr>
<td>broiled</td>
<td>hormone</td>
</tr>
<tr>
<td>seared</td>
<td>farmed</td>
</tr>
<tr>
<td>braised</td>
<td>hand picked</td>
</tr>
<tr>
<td>stirfried</td>
<td>Wild caught</td>
</tr>
<tr>
<td>flamebroiled</td>
<td>Farmhouse</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(d) MENUDESC: chicken</th>
<th>(f) MENUDESC:</th>
<th>(g) MENUDESC:</th>
</tr>
</thead>
<tbody>
<tr>
<td>slices</td>
<td>real mashed</td>
<td>crisp</td>
</tr>
<tr>
<td>bits</td>
<td>mashed</td>
<td>crispy bacon</td>
</tr>
<tr>
<td>cubes</td>
<td>creamy mashed</td>
<td>crispy bacon</td>
</tr>
<tr>
<td>pieces</td>
<td>smoked</td>
<td>crispy bacon</td>
</tr>
<tr>
<td>strips</td>
<td>smooth mashed</td>
<td>crispy bacon</td>
</tr>
<tr>
<td>chunks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>morsels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pcs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cuts</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(h) MENUDESC:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>roasted</td>
<td>roasted potatoes</td>
</tr>
<tr>
<td>roast</td>
<td>roasted potatoes</td>
</tr>
<tr>
<td>roasted chicken</td>
<td>roasted salmon</td>
</tr>
<tr>
<td>roast chicken</td>
<td>roasted salmon</td>
</tr>
<tr>
<td>roasted pork</td>
<td>roasted tomato</td>
</tr>
<tr>
<td>roasted pork</td>
<td>roasted tomato</td>
</tr>
</tbody>
</table>

chicken (e.g., wings, thighs, etc.) and the former more often used as a synonym for chunks.

Panes (f), (g), and (h) reveal price differences due to slight variations in word form. We find that, even though crispy has a higher weight than crisp, crisp bacon is more expensive than crispy bacon. We also find that food items prefixed with roast lead to more expensive prices than the similar roasted, except in the case of pork, though here the different forms may be evoking two different preparation styles.

Also of note is the slight difference between the nonstandard mash potato and mashed potato. We observed lower weights with other nonstandard spellings, notably portobella having lower weight than each of the more common spellings portabella, portobello, and portabello.

6 Restaurant Price Range Prediction

In addition to predicting the prices of individual menu items, we also considered the task of predicting the price range listed for each restaurant on its Yelp page. The values for this field are integers from 1 to 4 and indicate the price of a typical meal from the restaurant.

For this task, we again train an $\ell_1$-regularized linear regression model with integral price ranges as the true output values $y_i$. Each input $x_i$ corresponds to the feature vector for an entire restaurant. For evaluation, we round the predicted values to the nearest integer: $\hat{y}_i = \text{ROUND}(w^\top x_i)$ and report the corresponding mean absolute error and accuracy.

We compared this simple approach with an ordinal regression model (McCullagh, 1980) trained with the same $\ell_1$ regularizer and noted very little improvement (77.32% vs. 77.15% accuracy for METADATA). Therefore, we only report in this section results for the linear regression model.

In addition to the feature sets used for individual menu item price prediction, we used features on reviews (REVIEWS). Specifically, we used binary features for unigrams, bigrams, and trigrams in the full set of reviews for each restaurant. A stopword list was derived from the training data. Bigrams and trigrams were filtered if they ended with stopwords. Additionally, features occurring fewer than three times in the training set were discarded.

\*This list is included in the supplementary material.
Table 4: Results for restaurant price range prediction. MAE = mean absolute error, Acc = classification accuracy (%), nf = total number of features, nnz = number of features with non-zero weight.

<table>
<thead>
<tr>
<th>Features</th>
<th>MAE</th>
<th>Acc.</th>
<th>nf</th>
<th>nnz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict mode</td>
<td>0.5421</td>
<td>48.22</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>M*ENU</td>
<td>0.3875</td>
<td>66.29</td>
<td>1,910,622</td>
<td>995</td>
</tr>
<tr>
<td>METADATA</td>
<td>0.2372</td>
<td>77.15</td>
<td>591</td>
<td>219</td>
</tr>
<tr>
<td>REVIEWS</td>
<td>0.2172</td>
<td>79.76</td>
<td>3,027,470</td>
<td>1,567</td>
</tr>
<tr>
<td>+METADATA</td>
<td>0.2111</td>
<td>80.36</td>
<td>3,027,943</td>
<td>1,376</td>
</tr>
</tbody>
</table>

6.1 Results

Our results for price range prediction are shown in Table 4. Predicting the most frequent price range gave us an accuracy of 48.22%. Performance improvements were obtained by separately adding menu (M*ENU), metadata (METADATA), and review features (REVIEWS). Unlike individual item price prediction, the reviews were more helpful than the menu features for predicting overall price range. This is not surprising, since reviewers will often generally discuss price in their reviews. We combined metadata and review features to get our best accuracy, exceeding 80%.

We also wanted to perform an analysis of sentiment in the review text. To do this, we trained a logistic regression model predicting polarity for each review; we used the REVIEWS feature set, but this time considering each review as a single training instance. The polarity of a review was determined by whether or not its star rating was greater than the average rating across all reviews in the dataset (3.7 stars). We achieved an accuracy of 87% on the test data. We omit full details of these models because the polarity prediction task for user reviews is well-known in the sentiment analysis community and our model is not an innovation over prior work (Pang and Lee, 2008). However, our purpose in training the model was to use the learned weights for understanding the text in the reviews.

6.2 Interpreting Reviews

Given learned models for predicting a restaurant’s price range from its set of reviews as well as polarity for each review, we can turn the process around and use the feature weights to analyze the review text. Restricting our attention to reviews of 50–60 words, Table 5 shows sample reviews from our test set that lead to various predictions of price range and sentiment.\(^5\)

This technique can also be useful when trying to determine the “true” star rating for a review (if provided star ratings are noisy), or to show the most positive and most negative reviews for a product within a particular star rating. The 5-point scale is merely a coarse approximation to the reviewer’s mental state; using fitted models can provide additional clues to decode the reviewer’s sentiment.

We can also do a more fine-grained analysis of review text by noting the contribution to the price range prediction of each position in the text stream. This is straightforward because our features are simply \(n\)-grams of the review text. In Figure 2, we show the influence of each word in a review sentence on the predicted polarity (brown) and price range (yellow). The height of a bar at a given position is proportional to the sum of the feature weights for every unigram, bigram, and trigram containing the token at that position (there are at most 6 active \(n\)-grams at a position).

The first example shows the smooth shift in expressed sentiment from the beginning of the sentence to the end. The second sentence is a difficult example for sentiment analysis, since there are several positive words and phrases early but the sentiment is chiefly expressed in the final clause. Our model noted the steady positive sentiment early in the sentence but identified the crucial negation due to strong negative weight on bigrams \textit{fresh but, left me}, and \textit{me yearning}. In both examples, the yellow bars show that price estimates are reflected mainly through isolated mentions of offerings and amenities (\textit{drinks, atmosphere, security, good service}).

7 Joint Prediction of Price and Sentiment

Although we observe no interesting correlation (\(r = 0.06\)) between median star rating and median item price in our dataset, this does not imply that sen-

\(^5\)To choose the 9 reviews in the table, we took the reviews from our test set in the desired length range and computed predicted sentiment and price range for each; then we scaled the predicted price range so that its range matched that of predicted sentiment, and maximized various linear combinations of the two. This accounts for the four corners. The others were found by maximizing a linear combination of one (possibly negated) prediction minus the absolute value of the other.
great place to get fast food that tastes good. paneer and chicken are both good. i would prefer to go thursday thru saturday night. thats when they have their good shift working. also it stays open late until 4 am on weekends. really enjoyable!

great dark. sexy atmosphere. good service. nice variety of tasty cocktails. sushi tasted good. food was fresh but nothing left me yearning to return.

Table 5: Reviews from the test set deemed by our model to have particular values of sentiment and price.

Figure 2: Local (position-level) sentiment (brown) and price (yellow) estimates for two sentences in the test corpus.

ment and price are independent of each other.\textsuperscript{6} We try to capture this interaction by modeling at the same time review polarity and item price: we consider the task of jointly predicting aggregate sentiment and price for a restaurant.

For every restaurant in our dataset, we compute its median item price $\hat{p}$ and its median star rating $\hat{r}$. The average of these two values for the entire dataset ($\$8.69$ and $3.55$ stars) split the plane ($\vec{p}, \vec{r}$) in four sections: we assign each restaurant to one of these quadrants which we denote $\downarrow \ominus$, $\downarrow \oplus$, $\uparrow \ominus$, and $\uparrow \oplus$. This allows us to train a 4-class logistic regression model using the REVIEWS feature set for each restaurant. We achieve an accuracy of 65% on the test data, but we are mainly interested in interpreting the estimated feature weights.

7.1 Analysis

To visualize the top feature weights learned by the model, we have to map the four weight vectors learned by the model back to the underlying two-dimensional sentiment/price space. Therefore, we compute the following values:

\[
\begin{align*}
\vec{w}_\downarrow &= \vec{w}_{\uparrow \ominus} + \vec{w}_{\downarrow \ominus} - (\vec{w}_{\uparrow \ominus} + \vec{w}_{\downarrow \ominus}) \\
\vec{w}_\uparrow &= \vec{w}_{\uparrow \ominus} + \vec{w}_{\downarrow \ominus} - (\vec{w}_{\uparrow \ominus} + \vec{w}_{\downarrow \ominus})
\end{align*}
\]

\textsuperscript{6}Price and sentiment are both endogenous outcomes reflecting the characteristics of the restaurant. E.g., “better” restaurants can charge higher prices.
We then select for display the features which are the furthest from the origin ($\max w_5^2 + w_\star^2$) and represent the selected $n$-grams as points in the sentiment/price space to obtain Figure 3.

We notice that the spread of the sentiment values is larger, which suggests that reviews give stronger clues about consumer experience than about the cost of a typical meal. However, obvious price-related adjectives (inexpensive vs. expensive) appear in this limited selection, as well as certain phrases indicating both sentiment and price (overpriced vs. very reasonable). Other examples of note: gem is used in strongly-positive reviews of cheap restaurants; for expensive restaurants, reviewers use highly recommended or amazing. Also, phrases like no flavor and manager appear in negative reviews of more expensive restaurants, while dirty appears more often in negative reviews of cheaper restaurants.

8 Conclusion

We have explored linguistic relationships between food prices and customer sentiment through quantitative analysis of a large corpus of menus and reviews. We have also proposed visualization techniques to better understand what our models have learned and to see how they can be applied to new data. More broadly, this paper is an example of using extrinsic variables to drive model-building for linguistic data, and future work might explore richer extrinsic variables toward a goal of task-driven notions of semantics.

Acknowledgments

We thank Julie Baron, Ric Crabbe, David Garvett, Laura Gimpel, Chenxi Jiao, Elaine Lee, members of the ARK research group, and the anonymous reviewers for helpful comments that improved this paper. This research was supported in part by the NSF through CAREER grant IIS-1054319 and Sandia National Laboratories (fellowship to K. Gimpel).

References


Figure 3: Top 50 features from joint prediction of price and sentiment. The black circle is the origin. See text for details on how the coordinates for each feature were computed. Insets show enlargements of dense areas of the graph.

